International Journal of Multimedia and Ubiquitous Engineering
Vol.9, No.4 (2014), pp.43-50
http://dx.doi.org/10.14257/ijmue.2014.9.4.05

Interaural Time Difference Estimation Using Generalized Cross-correlation with Maximum Likelihood Weighting in Reverberant Environments

Ji Hun Park[1] and Seung Ho Choi[2,*]

[1]Visual Display R&D Office, Samsung Electronics, Gyeonggi-do 443-742, Korea
[2] Dept. of Electronic and IT Media Engineering
Seoul National University of Science and Technology, Seoul 139-743, Korea

** Corresponding author: shchoi@seoultech.ac.kr*

Abstract

In this paper, an interaural time difference (ITD) estimation method is proposed for binaural speech separation in reverberant environments. First, the auditory signals are represented in the time-frequency (T-F) domain, and the ITD for each T-F bin is then estimated using generalized cross-correlation (GCC) with a maximum likelihood (ML) weighting function. In particular, the ML weighting function is designed to reduce the reverberation effect. Then, a mask is estimated by comparing the estimated ITD with the ITD corresponding to the location of the pre-defined target speech source. Finally, the target speech is separated by applying the mask to the auditory signals. It is shown that the proposed ITD estimation method outperforms a conventional cross-correlation-based ITD estimation method under reverberant conditions in terms of the signal-to-noise ratio (SNR) and signal-to-distortion ratio (SDR) of the separated speech signals.

Keywords: *Interaural time difference, generalized cross-correlation, maximum likelihood weighting, binaural speech separation, reverberant environment*

1. Introduction

In the human auditory system, a desired signal can be localized and separated by the difference in the signal arrival time at each ear [1]. In a similar fashion, binaural speech separation approaches have attempted to separate sounds into target speech and noise according to the interaural time differences (ITDs) when the sounds are captured from two different microphones [2-4]. In other words, these approaches estimate the ITDs in all time-frequency (T-F) bins, and then estimate a mask by selecting the ITDs that are consistent with the ITD corresponding to the location of the target speech signal. It should be noted that the estimated mask indicates whether a T-F bin mainly includes the target speech signal or noise signals dominantly [5, 6]. Consequently, the target speech signal can be retrieved by applying the estimated mask to the microphone signal. Therefore, a correct estimate of the ITD is essential to obtain high-quality separation performance.

In general, ITDs can be easily estimated as time lags in which cross-correlation (CC) between the recorded signals at a pair of microphones is maximized [7-9]. Although the CC-based approach is very efficient and simple, certain problems may occur in reverberant environments. For example, echoes arise from room reverberation, which weakens the reliability of the ITD estimate and degrades the performance of the binaural speech separation [10-12]. Therefore, reverberations make binaural speech separation a very difficult task. To

ISSN: 1975-0080 IJMUE

increase the feasibility of a binaural speech separation system in reverberant environments, it is necessary to develop an ITD estimation method that is robust to reverberant conditions.

Thus, this paper proposes a reverberation-robust ITD estimation method. To this end, the ITD for each T-F bin is estimated using generalized CC (GCC) with a maximum likelihood (ML) weighting function, where the ML weighting function is designed to reduce the reverberant effects.

Following this introduction, Section 2 describes a binaural speech separation system employing the proposed ITD estimation method. Then, Section 3 proposes a robust ITD estimation method in reverberant environments by incorporating the GCC with an ML weighting function. Next, the performance of the proposed ITD estimation method is evaluated in Section 4. Finally, this paper is concluded in Section 5.

2. Binaural Speech Separation

Figure 1 shows a block diagram of a binaural speech separation system employing the proposed ITD estimation method. As shown in the figure, the system is mainly composed of four processing modules such as gammatone analysis, ITD estimation, mask estimation, and speech reconstruction. The following subsections describe each of these processing modules in detail.

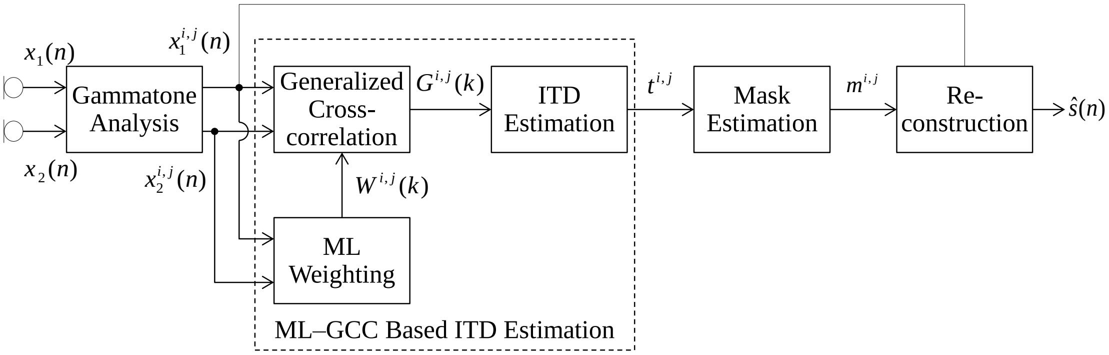

Figure 1. Block diagram of a binaural speech separation system using the proposed ML-GCC based ITD estimation method

2.1. Gammatone Analysis

Let $x_1(n)$ and $x_2(n)$ be dual-channel input signals sampled at a rate of 16 kHz. They are decomposed into auditory signals by a gammatone filterbank [13], which is a set of gammatone filters whose center frequencies are linearly spaced on an equivalent rectangular bandwidth (ERB) scale [14]. Note here that the gammatone filters have different group delays depending on the frequency bands, resulting in a phase shift between the frequency bands. To compensate for such a phase shift, a phase correction term is applied to the impulse response of each gammatone filter [15]. The decomposed auditory signals are then segmented at a frame rate of 100 Hz using a rectangular window with a time resolution of 20 ms. From now on, dual-channel auditory signals for the *ij*-th T-F bin are obtained and denoted as $x_1^{i,j}(n)$ and $x_2^{i,j}(n)$.

2.2. ITD Estimation

Basically, the ITD for each T-F bin is estimated as a time lag in which a CC of the corresponding T-F bin is maximized [9]. In this paper, to estimate the ITD, the auditory signal for the *ij*-th T-F bin is first de-trended to reduce the reverberation effects [16]. The de-trended auditory signals are then used to compute GCC instead of CC because GCC offers better performance than CC if a proper weighting function is used in the frequency domain [17, 18]. In addition, the weighting function is designed under a maximum likelihood criterion to further reduce the reverberation effect, and thus it is referred to as an ML weighting function. Consequently, the ITD estimation method in this paper is called an ML-GCC based method. Let us define the ITD for the *ij*-th T-F bin estimated using the proposed ML-GCC based method as $t^{i,j}$, which is further described in Section 3.

2.3. Mask Estimation

Assuming that the speech source is located in front of dual microphones, i.e., at an angle of 0°, the *ij*-th T-F bin is determined to be a speech bin if $t^{i,j}$ corresponds to an angle of 0°. Otherwise, it is determined to be a non-speech bin. Note that when the signal is arriving from a position at an angle of exactly 0°, there is theoretically no time difference between the two microphones, resulting in an ITD of 0. However, to allow some tolerance in the ITD, the region within ±10° is chosen as the location of the target speech signal. Therefore, a binary mask for the *ij*-th T-F bin, $m^{i,j}$, is given by

$$m^{i,j} = \begin{cases} 1, & \text{if } |ITD(i,j)| < \eta \\ \varepsilon, & \text{otherwise} \end{cases} \tag{1}$$

where ε represents a flooring factor and is set to 0.01. In addition, η is a pre-defined threshold that is calculated as

$$\eta = \left\lceil \frac{d \cdot \sin(\theta) \cdot f_s}{c} \right\rceil \tag{2}$$

where c is the speed of sound, d is the distance between the two microphones, f_s is the sampling frequency, and $\lceil \ \rceil$ indicates the ceiling operator. In addition, θ denotes the angle for the location of the sound source in radian and is set to $\pi/18$ as mentioned earlier. Thus, the value of η in (2) becomes 2.

2.4. Speech Reconstruction

The final step of the binaural speech separation system is to retrieve the separated target speech signal. To this end, auditory signals of the target speech are first estimated using multiplying masks to input the auditory signals. In particular, the mask in (1) is multiplied to $x_1^{i,j}(n)$. Because a phase shift between the frequency bands should be compensated, as described in Section 2.1, the estimated target speech for the *i*-th frame, $\hat{s}^i(n)$, is obtained by taking the sum of the auditory signals over all of the frequency bands as follows:

$$\hat{s}^i(n) = \sum_{j=0}^{J-1} x_1^{i,j}(n) \cdot m^{i,j} \tag{3}$$

where J is the number of frequency bands and is set to 32 in this paper. Finally, $\hat{s}(n)$ is reconstructed from $\hat{s}^i(n)$ using the overlap-and-add (OLA) method.

3. Proposed ML-GCC Based ITD Estimation

This section proposes a robust ITD estimation method robust in reverberant environments by incorporating the ML-GCC based approach for the ITD estimation. It is assumed here that the auditory signal in each T-F bin is composed of a direct signal and reverberant signals, where the reverberant signals are defined as down-scaled and delayed versions of the direct signal since they arrive later than the direct signal. Therefore, the auditory signal at the *ij*-th T-F segment can be represented as

$$\left|X_s^{i,j}(k)\right|^2 = \left|D_s^{i,j}(k)\right|^2 + \left|R_s^{i,j}(k)\right|^2 = \left|D_s^{i,j}(k)\right|^2 + g_s \cdot \left|D_s^{i,j}(k)\right|^2 \tag{4}$$

where $X_s^{i,j}(k)$ denotes the *k*-th frequency component of the auditory signal at the *ij*-th T-F bin, and $D_s^{i,j}(k)$ and $R_s^{i,j}(k)$ denote the direct and reverberant components of $X_s^{i,j}(k)$, respectively. In addition, $s(\in\{1,2\})$ indicates either of two channels, and g_s is a down-scaled gain for the *s*-th channel. From (4), the relationship between $X_s^{i,j}(k)$ and $R_s^{i,j}(k)$ is derived as

$$\left|R_s^{i,j}(k)\right|^2 = \frac{g_s}{1+g_s}\left|X_s^{i,j}(k)\right|^2. \tag{5}$$

By considering all the reflected signals as unwanted signals, an ML weighting function can be represented as

$$W^{i,j}(k) = \frac{\left|X_1^{i,j}(k)\right|\left|X_2^{i,j}(k)\right|}{\left|X_1^{i,j}(k)\right|^2\left|R_2^{i,j}(k)\right|^2 + \left|X_2^{i,j}(k)\right|^2\left|R_1^{i,j}(k)\right|^2} = \frac{1}{G\cdot\left|X_1^{i,j}(k)\right|\left|X_2^{i,j}(k)\right|} \tag{6}$$

where

$$G = \frac{g_1 g_2}{(1+g_1)(1+g_2)}. \tag{7}$$

In this paper, G is set to 0.45 from preliminary experiments. Using the ML weighting function defined in (6), an ML-GCC function is defined by [19]

$$G^{i,j}(k) = \sqrt{W^{i,j}(k)}\,X_1^{i,j}(k)\left(X_2^{i,j}(k)\right)^* \tag{8}$$

where * denotes a complex conjugate. The ITD for the *ij*-th T-F bin is then estimated as

$$ITD(i,j) = \arg\max_{\tau} C^{i,j}(\tau) \tag{9}$$

where $C^{i,j}(\tau)$ is the real part of the inverse Fourier transform of $G^{i,j}(k)$. In addition, τ is a time lag ranging from -8 to 8, corresponding to an angle of -90° to 90° at a sampling rate of 16 kHz.

4. Performance Evaluation

The performance of the proposed ML-GCC based ITD estimation method was evaluated in terms of the signal-to-noise ratio (SNR) and signal-to-distortion ratio (SDR) of the separated speech signals.

4.1. Database

In this experiment, the Computational Hearing in Multi-source Environments (CHiME) database [20] was used. The CHiME database was designed to model our everyday lives in acoustically cluttered indoor environments, such as a living room. In the database, 600 clean utterances spoken by 34 speakers were first reverberated using a room response. The room response was measured using a manikin in a real living room, which was placed at a fixed position 2 m away and directly in front of the origin of the target speech. Next, the noise signals were recorded in the same room, where the reverberation time, T_{60} , for the living room was 300 ms. Here, the noise signals included sounds from several sources generated in a typical living room, such as the voices of two adults and two children, a TV, kitchen and laundry appliances, footsteps, electronic gadgets, toys, pets, and outside noises. The reverberated target speech signals were then mixed with the noise signals by varying the SNR from -6 to 9 dB at a step of 3 dB.

4.2. SNR and SDR Measurements

To evaluate the proposed ITD estimation method, the SNR and SDR were measured to determine how much noise components were rejected and how much forbidden distortion and burbling artifacts were included in the processed signals, respectively [21]. It was possible to carry out the SNR and SDR evaluations because the CHiME database provided both clean utterances and noise signals. Assume that the estimated target speech, $\hat{s}(n)$, was a sum of the original clean signal, $c_T(n)$, and noise signal, $c_N(n)$, as [21]

$$\hat{s}(n) = c_T(n) + c_N(n) + c_A(n) \tag{10}$$

where $c_A(n)$ corresponded to the remaining artifacts after estimating the clean target speech. The SNR and SDR were then defined by computing the energy ratio, such that

$$SNR = 10\log_{10} \frac{\sum_n \left|c_T(n)\right|^2}{\sum_n \left|c_N(n)\right|^2} \tag{11}$$

and

$$SDR = 10\log_{10} \frac{\sum_n \left|c_T(n) + c_N(n)\right|^2}{\sum_n \left|c_A(n)\right|^2}\,. \tag{12}$$

Figure 2 compares the SNR and SDR of a binaural speech separation system using the proposed ML-GCC based ITD estimation method with those using a conventional CC-based ITD estimation method. In the figure, the SNRs and SDRs were averaged for all 600 utterances. As a reference, the SNRs measured from noisy speech utterances were displayed, which were illustrated by circular marks in Figure 2(a). For example, the SNR for a noisy speech mixed with noise at a 0 dB SNR was measured as -1.9 dB. As shown in Figure 2(a), a binaural speech separation system using the proposed ML-GCC based ITD estimation method yielded greater improvement in SNR than that using the conventional CC-based method under all SNR conditions. In particular, the proposed ML-GCC based ITD estimation method was more effective at lower SNRs.

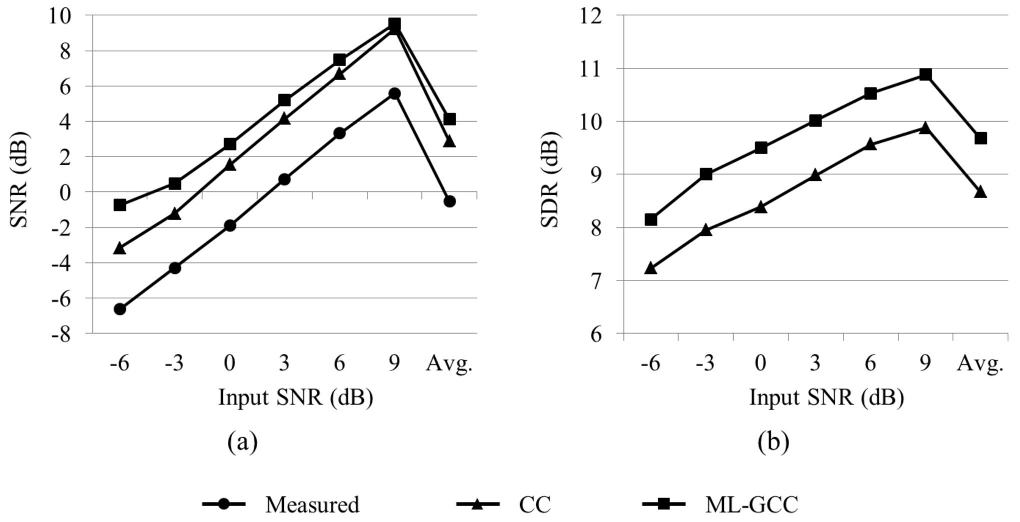

Figure 2. SNR and SDR comparisons of a binaural speech separation system using different ITD estimation methods under reverberant conditions: (a) SNR and (b) SDR measurements

Figure 2(b) also compares the SDR of a speech separation system using different ITD estimation methods. Note here that SDR values were compared only for two ITD estimation methods; the CC- and ML-GCC based methods because measuring the SDR for the original noisy speech utterances was meaningless. As shown in Figure 2(b), a binaural speech separation system using the proposed ML-GCC based ITD estimation method provided higher SDR than that using the conventional CC-based method for all SNR conditions. This implies that the proposed ML-GCC based ITD estimation method generated fewer burbling artifacts and distortion components than the conventional CC-based method did.

5. Conclusion

In this paper, a robust ITD estimation method was proposed for binaural speech separation in reverberant environments. To this end, an ITD in a particular frequency region for a given time frame was estimated using a GCC with an ML weighting function, where the ML weighting function was designed to reduce the effects of reverberation. To demonstrate the effectiveness of the proposed method, the SNR and SDR of separated speech signals obtained using a binaural speech separation system were compared when employing the proposed ML-GCC based ITD estimation method and a conventional CC-based ITD estimation method. It was shown from the comparison that the proposed ML-GCC based ITD estimation method outperformed the conventional CC-based method under reverberant conditions.

Acknowledgements

This study was (partially) supported by Seoul National University of Science and Technology.

References

[1] A. S. Bregman. Auditory Scene Analysis: the Perceptual Organization of Sound, MIT Press, Cambridge, MA, **(1999)**.

[2] D. L. Wang and G. J. Brown, "Computational Auditory Scene Analysis: Principle, Algorithms and Applications", IEEE Press, Piscataway, NJ, **(2006)**.

[3] S. M. Kim and H. K. Kim, "Hybrid Probabilistic Adaptation Mode Controller for Generalized Sidelobe Cancellers Applied to Multi-Microphone Speech Enhancement", Digital Signal Processing, vol. 25, **(2014)** February, pp. 123-133.

[4] S. M. Kim and H. K. Kim, "Probabilistic Spectral Gain Modification Applied to Beamformer-Based Noise Reduction in a Car Environment", IEEE Transactions on Consumer Electronics, vol. 57, no. 2, **(2011)** May, pp. 866-872.

[5] K. M. Jeon, N. I. Park, H. K. Kim, M. K. Choi and K. I. Hwang, "Mechanical Noise Suppression Based on Non-Negative Matrix Factorization and Multi-Band Spectral Subtraction for Digital Cameras", IEEE Transactions on Consumer Electronics, vol. 59, no. 2, **(2013)** May, pp. 296-302.

[6] S. M. Kim, H. K. Kim, S. J. Lee and Y. Lee, "Multiple Likelihood Ratio Test-Based Voice Activity Detection Robust to Impact Noise in a Car Environment", Information: an International Interdisciplinary Journal, vol. 16, no. 3B, **(2013)** March, pp. 3141-3151.

[7] K. J. Palomäki, G. J. Brown and D. L. Wang, "A Binaural Processor for Missing Data Speech Recognition in the Presence of Noise and Small-Room Reverberation", Speech Communication, vol. 43, no. 4, **(2004)** September, pp. 361-378.

[8] C. Kim, R. M. Stern, K. Eom and J. Lee, "Automatic Selection of Thresholds for Signal Separation Algorithms Based on Interaural Delay", Proceedings of Interspeech, Makuhari, Japan, **(2010)** September, pp. 729-732.

[9] J. H. Park, J. S. Yoon and H. K. Kim, "HMM-Based Mask Estimation for a Speech Recognition Front-End Using Computational Auditory Scene Analysis", IEICE Transactions on Information and Systems, vol. E91-D, no. 9, **(2008)** September, pp. 2360-2364.

[10] C. J. Darwin and R. W. Hukin, "Effects of Reverberation on Spatial, Prosodic, and Vocal-Tract Size Cues to Selective Attention", Journal of the Acoustical Society of America, vol. 108, no. 1, **(2000)** July, pp. 335-342.

[11] Y. R. Oh and H. K. Kim, "A Hybrid Acoustic and Pronunciation Model Adaptation Approach for Non-Native Speech Recognition", IEICE Transactions on Information and Systems, vol. E93-D, no. 9, **(2010)** September, pp. 2379-2387.

[12] J. A. Kang and H. K. Kim, "Adaptive Redundant Speech Transmission over Wireless Multimedia Sensor Networks Based on Estimation of Perceived Speech Quality", Sensors, vol. 11, no. 9, **(2011)** September, pp. 8469-8484.

[13] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth and P. Rice, "An Efficient Auditory Filterbank Based on the Gammatone Functions", Applied Psychology Unit, Report 2341, Cambridge, U.K., **(1988)**.

[14] B. R. Glasberg and B. C. J. Moore, "Derivation of Auditory Filter Shapes from Notched-Noise Data", Hearing Research, vol. 47, no. 1-2, **(1990)** August, pp. 103-138.

[15] M. Cooke, "Modelling Auditory Processing and Organization", Cambridge University Press, Cambridge, U.K., **(2005)**.

[16] B. Yegnanarayana and P. Satyanarayana, "Enhancement of Reverberant Speech Using LP Residual Signal", IEEE Transactions on Speech and Audio Processing, vol. 8, no. 3, **(2000)** May, pp. 267-281.

[17] P. J. Hahn, V. J. Mathews and T. D. Tran, "Adaptive Realization of a Maximum Likelihood Time Delay Estimator", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Atlanta, GA, **(1996)** May, pp. 3121- 3124.

[18] N. I. Park and H. K. Kim, "Artificial Bandwidth Extension of Narrowband Speech Applied to CELP-Type Speech Coding", Information: an International Interdisciplinary Journal, vol. 16, no. 3B, **(2013)** March, pp. 3153-3164.

[19] C. H. Knapp and G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 24, no. 4, **(1976)** August, pp. 320-327.

[20] H. Christensen, J. Barker, N. Ma and P. Green, "The CHiME Corpus: a Resource and a Challenge for Computational Hearing in Multisource Environments", Proceedings of Interspeech, Makuhari, Japan, **(2010)** September, pp. 1918-1921.

[21] E. Vincent, R. Gribonval and C. Févotte, "Performance Measurement in Blind Audio Source Separation", IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, **(2006)** July, pp. 1462-1469.

Authors

Ji Hun Park

He received a B.S. degree in Electronics Engineering from Kwangwoon University, Korea in 2006. He then received both M.S. and Ph.D. degrees in Information and Communications Engineering from the Gwangju Institute of Science and Technology (GIST), Korea in 2008 and 2013, respectively. Since October 2013, he has been a senior engineer at Samsung Electronics, Co. Ltd., Korea. His current research interests include speech recognition and keyword spotting.

Seung Ho Choi

He received a B.S. degree in Electronic Engineering from Hanyang University, Korea in 1991. He then received both M.S. and Ph.D. degrees in Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST), Korea in 1993 and 1999, respectively. He was a senior researcher at Samsung Advanced Institute of Technology, Korea, from 1996 to 2002. He was a visiting professor at University of Florida, USA, from 2008 to 2009. Since August 2002, he has been with the Department of Electronic and IT Media Engineering at Seoul National University of Science and Technology as a professor. His current research interests include speech and audio coding, acoustic signal processing, speech recognition.